

Inverting the signature of a path

Terry J. Lyons
University of Oxford

Weijun Xu
University of Warwick

July 2, 2015

Abstract

The aim of this article is to develop an explicit procedure that enables one to reconstruct any \mathcal{C}^1 path (when parametrized at uniform speed) from its signature. A key ingredient in the construction is the use of symmetrization that separates the behavior of the path at small and large scales. We also give detailed stability properties of our inversion procedure.

Key words: signature, inversion, symmetrization.

AMS Classification: 70.

1 Introduction

Paths arise naturally in the study of time evolving systems. The signature of a path in \mathbb{R}^d is a formal power series with d indeterminates whose coefficients are iterated integrals of the path. Since it is the main object of study in this article, we give a precise definition below.

Definition 1.1. Let $\gamma : [0, T] \rightarrow \mathbb{R}^d$ be a path of finite length, and $\{e_1, \dots, e_d\}$ be the standard basis of \mathbb{R}^d . For any integer n and word $w = e_{i_1} \cdots e_{i_n}$, define

$$C_\gamma(w) = \int_{0 < u_1 < \cdots < u_n < T} d\gamma_{u_1}^{i_1} \cdots d\gamma_{u_n}^{i_n}.$$

The signature of γ is then the formal series

$$X(\gamma) = \sum_{n=0}^{+\infty} \sum_{|w|=n} C_\gamma(w)w,$$

where the second sum is taken over all words w with length n , and we use the convention that $|w| = 0$ denotes the empty word.

The signature is a definite integral over a fixed time interval where γ is defined. Re-parametrizing γ does not change its signature. One reason to look at the signatures is that they contain important information about the paths. For example, the first level coefficients $\{C_\gamma(w) : |w| = 1\}$ reproduce the increment of the path,

and the second level collection $\{C_\gamma(w) : |w| = 2\}$ represents the area enclosed by the projection of the path on the $e_i - e_j$ planes, etc.

The study of these iterated integrals dates back to K.T.Chen in 1950's. In a series of papers ([Che54], [Che57], [Che58]), he showed that the map $\gamma \mapsto X(\gamma)$ is a homomorphism from the monoid of paths with concatenation to the tensor algebra over \mathbb{R}^d , and proved that two piecewise regular paths have the same signatures if and only if they differ by a re-parametrization. He further developed these iterated integrals for paths on manifolds ([Che77]), and used them to relate analysis on a manifold to the homology of its path spaces.

Recently, Hambly and the first author quantified and extended Chen's uniqueness of signature result to all bounded variation paths. It was shown in [HL10] that paths of finite length are uniquely determined by their signatures up to tree-like equivalence¹. A consequence of this result is that among all finite length paths with the same signature, there is a unique one with minimal length, called the tree-reduced path. It is then natural to ask the following question.

Problem 1.2. *Given the signature X , how does one reconstruct the tree-reduced path from it?*

Remark 1.3. In the case $d = 1$, the signature X simply takes the form $X = \exp(ax)$, and the tree-reduced path is simply the straight line from 0 to a on \mathbb{R} . The problem is interesting only when $d \geq 2$, where evolutions of the path are in general non-commutative.

The proofs of the results mentioned above are all pure uniqueness arguments, and they do not give clues on how the path might be recovered from the signature.

A naive approach to Problem 1.2 can be to try to reproduce the path from the interpretation of each term in the signature, as the explanation of the meanings of $\{C_\gamma(w) : |w| \leq 2\}$ above. However, these intuitive interpretations break down when the level $|w| = n$ gets large, and it is essentially impossible to proceed this way to recover much finer information of the path beyond the increment and area. Thus, certain operations on signatures that can reveal local information of the path would be necessary for the reconstruction.

There have been recent attempts to Problem 1.2. In establishing the uniqueness of signature for Brownian motion sample paths, Le Jan and Qian ([LJQ13]) constructed polygonal approximations to the Brownian paths by using the information in the signatures only. This approximation scheme has also been extended to diffusions ([GQ15]) and Gaussian processes ([BG14]). However, in all these situations, the constructions are inexplicit and, even for the approximation at any fixed scale, it requires the use of the *whole signature sequence* rather than any of its truncations. Thus, it is hard to turn this scheme into an effective algorithm.

¹As defined in [HL08] and [HL10], a path $\gamma : [0, T] \rightarrow \mathbb{R}^d$ is tree-like if there exists a positive continuous function h defined on $[0, T]$ such that $h(0) = h(1) = 0$ and

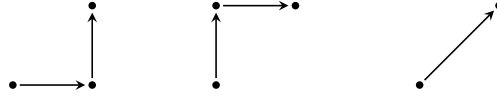
$$|\gamma_s - \gamma_t| \leq h(s) + h(t) - 2 \inf_{u \in [s, t]} h(u)$$

for all $s < t \in [0, T]$. Heuristically, one can think of a tree-like paths as being a null-path as a control; the trajectories are completely canceled out by themselves. Two paths $\gamma, \tilde{\gamma}$ are tree-equivalent if $\gamma * \tilde{\gamma}^{-1}$ is tree-like.

In the recent article [LX15], by using a construction from hyperbolic geometry, the authors gave an explicit inversion scheme for piecewise linear paths, together with its stability properties. But this scheme makes an essential use of the piecewise linearity of the path, and it is not clear whether it could be extended to more general situations.

1.1 Main result and the strategy

The main goal of this article is to develop a reconstruction procedure to Problem 1.2 for all \mathcal{C}^1 paths (when parametrized at uniform speed). As mentioned in Remark 1.3, the key to establish an effective reconstruction algorithm is to recover the non-commutative evolution of the path in the correct order: the following three paths in \mathbb{R}^2 have the same increments. Their signatures agree on level $n = 1$, but start to differ when $n \geq 2$.



In fact, if we let x and y denote the standard basis of \mathbb{R}^2 , then the signatures of these three paths are the formal series $e^x e^y$, $e^y e^x$ and e^{x+y} , respectively.

As one can see, the order of the evolution of the paths are captured in the signatures through the ordered letters that consist of the words w 's. Our main result is that, by symmetrizing the signatures, we can average out the non-commutativity of the path at small scales but still preserve the order of its evolution at larger scales. Thus, for any given integer k , we produce a piecewise linear path from the signature that approximates the original path at scale $\frac{1}{k}$. The main theorem could be loosely stated as follows.

Theorem 1.4. *Let γ be a \mathcal{C}^1 path in \mathbb{R}^d (when parametrized at uniform speed) with signature $X = \{C(w) : |w| \geq 0\}$. For any k , by using the information in the signature up to level $|w| = k^2 \log k$, we construct a piecewise linear path $\tilde{\gamma}$ with k linear pieces such that*

$$\|\tilde{\gamma} - \gamma\|_{\mathcal{C}^1} < C\eta_k$$

when both γ and $\tilde{\gamma}$ are parametrized at uniform speed in the same time interval, and $\eta_k \rightarrow 0$ with rate depending on the modulus of continuity of $\dot{\gamma}$.

Remark 1.5. The dependence of η_k on the regularity of $\dot{\gamma}$ will be made explicit below. In the case $\gamma \in \mathcal{C}^{1+\alpha}$, we have $\eta_k = \mathcal{O}(k^{-\frac{\alpha}{2}})$ (see Remark 4.1 below). But we do not expect this rate to be optimal (nor is the level $k^2 \log k$), and an improvement is possible at least in the case when $\alpha < 1$.

Remark 1.6. The \mathcal{C}^1 assumption is mainly for technical convenience. We expect that the same reconstruction procedure with slight modifications works for general finite length paths, but a much more technically involved argument would be needed for the proof.

1.2 Some notations and assumptions

Since we will fix the path γ in the rest of the article, we simply write X instead of $X(\gamma)$ for its signature. We will also use ℓ^1 norm on \mathbb{R}^d throughout the article. The reason for this is that symmetrization has a nice behaviour under ℓ^1 norm rather than the usual Euclidean norm. Thus, the length of γ is

$$|\gamma| = |\gamma|_{\ell^1} = \sum_{j=1}^d \int_0^T |\gamma_t^{(j)}| dt.$$

Since the signature is invariant under parametrization, we always assume γ is parametrized on the unit interval $[0, 1]$ with uniform speed such that

$$|\dot{\gamma}| \equiv L. \quad (1.1)$$

Also, the standing assumption $\gamma \in \mathcal{C}^1$ should be understood that γ is \mathcal{C}^1 when parametrized as (1.1). Since all the quantitative estimates below will depend on the continuity properties of the derivative $\dot{\gamma}$, we let

$$\delta_k = \sup_{|s-t| < \frac{1}{k}} |\dot{\gamma}_s - \dot{\gamma}_t| \quad (1.2)$$

denote its modulus of continuity.

It is sometimes convenient to write X in terms of tensors, so we let

$$X^n = \sum_{|w|=n} C(w)w = \int d\gamma_{u_1} \otimes \cdots \otimes d\gamma_{u_n}$$

be the n -th level signature, and thus $X = \sum_{n \geq 0} X^n$, an element with value in the free tensor algebra over \mathbb{R}^d , denoted by $T(\mathbb{R}^d)$.

Finally, since there is no essential difference between two and higher dimensions, for notational simplicity, we will focus on the two dimensional case only. Section 4.5 explains how the procedure extends to higher dimensions.

1.3 Extensions and further questions

Theorem 1.4 is based on the assumption that X is the signature of a \mathcal{C}^1 path. It is however not clear a priori what kind of elements in the tensor algebra do come from such a path. The following question then becomes natural.

Question 1.7. *Identify the elements in the tensor algebra that are the signatures of some paths.*

It is clear that X should be a group-like element in $T(\mathbb{R}^d)$ in order for it to be a signature, but this condition is not sufficient. One possible approach to the above question is to go through the reconstruction procedure for the tensor element X to see whether it indeed produces an authentic path. However, such an argument would inevitably be very complicated, and we do not expect it to give an intrinsic description of the structure for the signature. An intrinsic criterion on the tensor element X itself would be much more desirable.

Recently, using the definition of signature in [Lyo98] for rough paths, [BGLY14] extended the uniqueness result in [HL10] to geometric rough paths that may not have finite length. Thus, another natural question is to develop an inversion procedure for those paths:

Question 1.8. *Assuming $X \in T(\mathbb{R}^d)$ is the signature of some (possibly rough) path, how could one reconstruct that path from X ?*

We should point out that the procedure developed in this article does depend crucially on the assumption that γ has finite length. In fact, each line segment in the piecewise linear path $\tilde{\gamma}$ in Theorem 1.4 will approximate their counter parts in the original path γ with respect to ℓ^1 length. Thus, the development of the inversion procedure for rough paths would require new treatments.

1.4 Structure of the article

The rest of the article is structured as follows. In Section 2, we set up the symmetrization procedure for signatures. Section 3 is devoted to the proof of the concentration property of the symmetrized signatures with quantitative estimates. It contains mostly technical statements. Finally, in Section 4, we give a detailed description of the reconstruction procedure of the path from the symmetrized signatures. The concentration property proved in Section 3 would be essential to guarantee the validity and stability of the reconstruction.

Acknowledgements

The research of Terry Lyons is supported by EPSRC grant EP/H000100/1 and the European Research Council under the European Unions Seventh Framework Program (FP7-IDEAS-ERC) / ERC grant agreement nr. 291244. Terry Lyons acknowledges the support of the Oxford-Man Institute. Weijun Xu has been supported by the Oxford-Man Institute through a scholarship during his time as a student at Oxford. He is now supported by Leverhulme trust.

2 Symmetrization

Let X be the signature of an unknown C^1 path γ in \mathbb{R}^2 and k be a fixed large number. Our aim is to construct from X a piecewise linear path

$$\tilde{\gamma} = \tilde{\gamma}_1 * \cdots * \tilde{\gamma}_k,$$

where each $\tilde{\gamma}_j$ is a line segment of the form

$$\tilde{\gamma}_j = \frac{\tilde{L}}{k} (a_x^{(j)} \rho_j x + a_y^{(j)} (1 - \rho_j) y).$$

Here, $\rho_j \in [0, 1]$ represents the unsigned direction of this linear piece, $a_x^{(j)}, a_y^{(j)} \in \{\pm 1\}$ reflect the signs of the two directions, and \tilde{L} is an approximation to L , the ℓ^1 length of γ . In the rest of the article, by merely using the knowledge of X , we will produce these parameters such that

$$\sup_{0 \leq j \leq k} \sup_{u \in [\frac{j-1}{k}, \frac{j}{k}]} \left| \tilde{L} (a_x^{(j)} \rho_j, a_y^{(j)} (1 - \rho_j) - \dot{\gamma}_u) \right| < C \eta_k$$

for some constant C depending on the path γ only, and $\eta_k \rightarrow 0$ depending on the modulus of continuity of $\dot{\gamma}$. This immediately implies that the reconstructed path $\tilde{\gamma}$ is close to γ in \mathcal{C}^1 norm when parametrized properly.

As mentioned in the introduction, the main strategy to get such a piecewise linear approximation is to use symmetrization to average out the non-commutativity at small scales while still keeping the order at large scales. In order to get a rough idea how the procedure works, we first briefly recall from [LX15] the inversion scheme for integer lattice paths. This can be decomposed into two steps:

1. Identify the unique square free word² w such that $C(w) \neq 0$. The order of the letters in w gives the directions (up to the sign) of each piece of the lattice paths.
2. Move one level up in the signature to recover the sign as well as the length of each piece.

At first glance, this procedure seems to crucially depend on the very special structure of lattice paths, and does not generalize directly to other situations. In particular, the vanishing/non-vanishing property of coefficients of square free words does not carry over to more general cases where the path can move along any direction in the plane. But fortunately, it turns out that similar results still hold if we replace the strict zero/non-zero criterion by a more robust notion of degeneracy/non-degeneracy.

We now give a simple example to illustrate how we can combine symmetrization and the new non-degeneracy notion to recover the directions.

Example 2.1. Let X be the signature of some bounded variation path γ , and we would like to recover from X the increments

$$\Delta x := x_1 - x_0, \quad \text{and} \quad \Delta y := y_1 - y_0.$$

One could of course get the exact values of this pair directly from the first level signature X^1 . The symmetrization method given below is more complicated, but has the advantage that it applies to general situations.

For each n and $0 \leq \ell \leq n$, we let

$$\mathcal{S}^n(\ell) = \sum C(w),$$

where the sum is taken over all words with length $|w| = n$ that contain ℓ x 's and $(n - \ell)$ y 's. Thus, $\mathcal{S}^n(\ell)$ has the expression

$$\mathcal{S}^n(\ell) = \binom{n}{\ell} (\Delta x)^\ell (\Delta y)^{n-\ell}.$$

Note that for each n and ℓ , the left hand side above is the information available to us (from the signature), and the right hand side is its expression. It is standard that for fixed large n , the quantity $|\mathcal{S}^n(\ell)|$ is maximized near the value ℓ^* such that

$$\frac{\ell^*}{n - \ell^*} \approx \frac{|\Delta x|}{|\Delta y|}.$$

We could thus asymptotically recover the ratio $|\Delta x| : |\Delta y|$ by finding the maximizer ℓ^* of $\mathcal{S}^n(\ell)$, and this gives us the *unsigned direction* of the increment.

To recover the signs of Δx and Δy , one repeats the same trick as in integer lattice paths: moving one level up and comparing the signs. Finally, the magnitude of the increment could be obtained by a simple scaling.

²A word $w = e_{i_1} \cdots e_{i_n}$ is square free if for all $j = 1, \dots, n - 1$, we have $i_j \neq i_{j+1}$.

The example above illustrates the trivial case when $k = 1$. In order to recover more refined information of the path (for large k), instead of symmetrizing the whole signature, we divide high level signatures into k equal blocks, and symmetrize each block.

We let Δ_{k-1} denote the standard simplex

$$\Delta_{k-1} = \{0 < u_1 < \cdots < u_{k-1} < 1\},$$

and use \mathbf{u} denote the point $\mathbf{u} = (u_1, \dots, u_{k-1}) \in \Delta_{k-1}$. For each $\mathbf{u} \in \Delta_{k-1}$, we let

$$\Delta_{u_j} x = x_{u_j} - x_{u_{j-1}}, \quad \Delta_{u_j} y = y_{u_j} - y_{u_{j-1}}$$

denote the increments in relevant directions in the time interval $[u_{j-1}, u_j]$, and

$$|\Delta_{u_j} \gamma| = |\Delta_{u_j} x| + |\Delta_{u_j} y| \quad (2.1)$$

be the magnitude of the increments, which reflects the ℓ^1 norm we are working with. Similarly, we denote the increments of the j -th piece under standard uniform partition by

$$\Delta_j x = x_{j/k} - x_{(j-1)/k}, \quad \Delta_j y = y_{j/k} - y_{(j-1)/k},$$

and the same for $|\Delta_j \gamma|$.

If w is a word, we let $|w(x)|$ denote the number of letters x in w , and $|w(y)|$ denote the number of letters y . For any word $w = e_{i_1} * \cdots * e_{i_{k-1}}$ and multi-index $\ell = \{\ell_1, \dots, \ell_k\}$ with $0 \leq \ell_j \leq n$, we let $\mathcal{W}_k^{2n}(w, \ell)$ be the set of words

$$\mathcal{W}_k^{2n}(w, \ell) = \left\{ w' = w_1 * e_{i_1} * \cdots * e_{i_{k-1}} * w_k : |w_j(x)| = 2\ell_j, |w_j(y)| = 2n - 2\ell_j \right\}.$$

A typical word $w' \in \mathcal{W}_k^{2n}(w, \ell)$ where $w = e_{i_1} \cdots e_{i_{k-1}}$ has the form

$$\underbrace{****}_{w_1} e_{i_1} \underbrace{****}_{w_2} e_{i_2} \cdots e_{i_{k-2}} \underbrace{****}_{w_{k-1}} e_{i_{k-1}} \underbrace{****}_{w_k}.$$

Here, each w_j is a sub-word of length $2n$ with $2\ell_j$ letters x and $2n - 2\ell_j$ letters y . The two consecutive sub-words (blocks) w_{j-1} and w_j are separated by the letter e_{i_j} from w . For example, for $n = 2$ and $k = 1$, we have

$$\mathcal{W}_1^4(x, (1, 0)) = \{xyxy, yxyx\}, \quad \mathcal{W}_1^4(x, (1, 2)) = \{xyxx, yxxx\}$$

and

$$\mathcal{W}_1^4(y, (1, 1)) = \{xyxy, xyxy, yxyx, yxyx\}.$$

With this definition, we introduce the *symmetrized signatures*

$$\mathcal{S}_k^{2n}(w, \ell) := \sum_{w' \in \mathcal{W}_k^{2n}(w, \ell)} C(w'). \quad (2.2)$$

It is not hard to check that this quantity can be expressed by

$$\mathcal{S}_k^{2n}(w, \ell) = \int_{\Delta_{k-1}} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \prod_{j=1}^k \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} d\mathbf{u}. \quad (2.3)$$

We will see below that these $\mathcal{S}_k^{2n}(w, \ell)$'s are the only quantities we are going to use to recover the unsigned directions ρ_j 's. It is the recovery of the sign of each direction that requires extra information in the signature other than the \mathcal{S}_k^{2n} 's, and we will introduce those new quantities only when it becomes necessary.

Remark 2.2. We emphasize that (2.2) is the definition of the symmetrized signature; this information is available to us from the signature X . On the other hand, (2.3) is an expression of this quantity, and we will make use of this expression later to prove a priori bounds of the symmetrized signature.

Remark 2.3. The reason why we insert a letter e_{i_j} between every two consecutive symmetrized blocks is to let $\mathcal{S}_k^{2n}(w, \ell)$ have a closed form expression as in (2.3). This is mainly for technical convenience, and we expect results in the next section still hold true when the symmetrization is taken without using these e_{i_j} 's to separate blocks. Also, the symmetrization is taken only over even numbers of x 's and y 's in each block. This is to avoid cancellations of different signs inside the integration on the right hand side of (2.3).

3 Concentration of symmetrized signatures

The aim of this section is to prove a concentration property of the integral

$$\int_{\Delta_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^n d\mathbf{u} \quad (3.1)$$

for large k and n . This integral arises naturally by summing over all possible indices ℓ on the right hand side of (2.3), where the integrand then becomes

$$\frac{1}{2^k} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \prod_{j=1}^k ((\Delta_{u_j} x + \Delta_{u_j} y)^{2n} + (\Delta_{u_j} x - \Delta_{u_j} y)^{2n}) \sim C_k \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n}.$$

This concentration property roughly states that although the integration is taken over the whole simplex, when n is large, almost all its contribution comes from a very small subset of Δ_{k-1} . In fact, the domain of concentration is around the points $\mathbf{u} \in \Delta_{k-1}$ such that the product $\prod_j |\Delta_{u_j} \gamma|$ is maximized, and these maximizers cannot be far away from the 'standard' locations $\{\frac{j}{k}\}_{j=1}^{k-1}$. A quantitative statement will be given in Proposition 3.3 below. We first give a few useful lemmas.

Lemma 3.1. *For all large enough k and all $j = 1, \dots, k$, we have*

$$\frac{L - \delta_k}{k} \leq |\Delta_j \gamma| \leq \frac{L}{k}.$$

Proof. The inequality $|\Delta_j \gamma| \leq \frac{L}{k}$ follows immediately from the assumption that γ is parametrized at uniform speed.

For the lower bound, we let $I_j = [\frac{j-1}{k}, \frac{j}{k}]$. If both \dot{x}_u and \dot{y}_u keep their signs unchanged in the interval I_j , then we have

$$\left| \int_{I_j} \dot{x}_u du \right| = \int_{I_j} |\dot{x}_u| du, \quad \left| \int_{I_j} \dot{y}_u du \right| = \int_{I_j} |\dot{y}_u| du,$$

and it follows immediately that $|\Delta_j \gamma| = \frac{L}{k}$. If not, then either \dot{x} or \dot{y} is 0 at some point in the interval I_j . We suppose $\dot{y}_u = 0$ for some $u \in I_j$, then the continuity of $\dot{\gamma}$ implies that

$$\sup_{u \in I_j} |\dot{y}_u| \leq \delta_k,$$

and thus

$$|\dot{x}_u| \geq L - \delta_k$$

for all $u \in I_j$. In addition, the continuity of \dot{x} also implies that \dot{x}_u does not change its sign in I_j if k is large enough, so we have

$$|\Delta_j \gamma| \geq \left| \int_{I_j} \dot{x}_u du \right| = \int_{I_j} |\dot{x}_u| du \geq \frac{L - \delta_k}{k}.$$

This finishes the proof of the lemma. \square

The above lemma implies that at the equal partition points $\mathbf{u} = \{\frac{j}{k}\}$, the product $\prod_j |\Delta_j \gamma|$ is close to its largest possible value. On the other hand, if \mathbf{u} is far away from the standard location $\{\frac{j}{k}\}$, then the product $\prod_j |\Delta_{u_j} \gamma|$ must be small. This is the content of the next lemma.

Lemma 3.2. *Let k be fixed. If $|u_j - \frac{j}{k}| > \sqrt{\delta_k/L} + \sqrt{1/k}$ for some $0 \leq j \leq k$, then we must have*

$$\prod_{i=1}^k \left(|\Delta_{u_i} \gamma| / |\Delta_i \gamma| \right) < \left(1 - \frac{\delta_k}{L} - \frac{1}{k} \right)^k < \frac{1}{e}. \quad (3.2)$$

Proof. Since both sides of (3.2) are invariant under rescaling of length, we can assume without loss of generality that $L = 1$. Suppose

$$u_j - \frac{j}{k} = \epsilon$$

for some j and some ϵ , then $u_j = \frac{j}{k} + \epsilon$, and the sum of all increments before and after the time $t = u_j$ satisfy

$$\sum_{i=1}^j |\Delta_{u_i} \gamma| \leq \frac{j}{k} + \epsilon, \quad \sum_{i=j+1}^k |\Delta_{u_i} \gamma| \leq \frac{k-j}{k} - \epsilon. \quad (3.3)$$

Note that here we do not require ϵ to be positive. By the bound (3.3), the best possible maximum one can hope for $\prod_j |\Delta_{u_j} \gamma|$ is the case when we have

$$|\Delta_{u_i} \gamma| = \frac{1}{k} + \frac{\epsilon}{j}, \quad \forall i \leq j \quad \text{and} \quad |\Delta_{u_i} \gamma| = \frac{1}{k} - \frac{\epsilon}{k-j}, \quad \forall i \geq j+1,$$

which gives

$$\prod_{i=1}^k |\Delta_{u_i} \gamma| \leq \left(\frac{1}{k} + \frac{\epsilon}{j} \right)^j \cdot \left(\frac{1}{k} - \frac{\epsilon}{k-j} \right)^{k-j}.$$

Using Lemma 3.1, we get

$$\prod_{i=1}^k \left(|\Delta_{u_i} \gamma| / |\Delta_i \gamma| \right) \leq \left(\frac{(1+p\epsilon)^{\frac{1}{p}} (1-q\epsilon)^{\frac{1}{q}}}{1-\delta_k} \right)^k,$$

where $p = \frac{k}{j}$ and $q = \frac{k}{k-j}$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Now, let

$$f(x) = (1+px)^{\frac{1}{p}} (1-qx)^{\frac{1}{q}},$$

we have

$$f(0) = 1, \quad f'(0) = 0, \quad f''(0) = -(p+q) \leq -4.$$

It is then clear that if $\epsilon^2 > \delta_k + \frac{1}{k}$, we will have

$$f(\epsilon) < 1 - \delta_k - \frac{1}{k},$$

which in turn implies

$$\prod_{i=1}^k \left(|\Delta_{u_i} \gamma| / |\Delta_i \gamma| \right) < (1 - \delta_k - \frac{1}{k})^k < \frac{1}{e}.$$

To get the case for general L , one simply replaces δ_k by $\frac{\delta_k}{L}$. This finishes the proof of the lemma. \square

Now for large k , we let

$$\epsilon_k := \sqrt{\delta_k/L} + \sqrt{1/k}, \quad (3.4)$$

and also let E_{k-1} be the set

$$E_{k-1} = \{(u_1, \dots, u_{k-1}) : |u_j - \frac{j}{k}| < \epsilon_k, j = 1, \dots, k-1\}.$$

The main reason to add the quantity $\sqrt{1/k}$ in the definition of ϵ_k is to guarantee that Lemma 3.2 is still true even if $\delta_k = 0$. In general, since any smooth path that is not a straight line would have $\delta_k > \frac{C}{k}$ for all large k , this additional term would not affect the size of ϵ_k or E_{k-1} . It is now very easy to prove the following concentration property.

Proposition 3.3. *There exists $c > 0$ such that for all large enough k and $n = k^2 \log k$, we have*

$$\int_{\Delta_{k-1} \cap E_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^n d\mathbf{u} \geq (1 - e^{-cn}) \int_{\Delta_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^n d\mathbf{u}.$$

Proof. We let \mathcal{E}_{k-1} denote the set

$$\mathcal{E}_{k-1} = \{\mathbf{v} : |v_j - \frac{j}{k}| < \frac{1}{4k^2}, j = 1, \dots, k\}.$$

By Lemma 3.1 and the continuity of $\dot{\gamma}$, for k large enough and all $\mathbf{v}_j \in \mathcal{E}_{k-1}$, we have

$$\prod_{j=1}^k |\Delta_{v_j} \gamma| \geq e^{-\frac{1}{2}} \prod_{j=1}^k |\Delta_j \gamma|$$

Now, combining this with Lemma 3.2, and raising both to the power n , we get

$$\prod_{j=1}^k |\Delta_{u_j} \gamma|^n \leq e^{-\frac{n}{2}} \prod_{j=1}^k |\Delta_{v_j} \gamma|^n,$$

for all $\mathbf{u} \in \Delta_{k-1} \cap E_{k-1}^c$, $\mathbf{v} \in \Delta_{k-1} \cap \mathcal{E}_{k-1}$, and all n . Averaging both sides of the above in their respective domains, we get

$$\frac{1}{|\Delta_{k-1} \cap E_{k-1}^c|} \int_{\Delta_{k-1} \cap E_{k-1}^c} \prod_{j=1}^k |\Delta_{u_j} \gamma|^n d\mathbf{u} \leq \frac{e^{-\frac{n}{2}}}{|\Delta_{k-1} \cap \mathcal{E}_{k-1}|} \int_{\Delta_{k-1} \cap \mathcal{E}_{k-1}} \prod_{j=1}^k |\Delta_{v_j} \gamma|^n d\mathbf{v},$$

which in turn gives

$$\int_{\Delta_{k-1} \cap E_{k-1}^c} \prod_{j=1}^k |\Delta_{u_j} \gamma|^n d\mathbf{u} \leq C_k e^{-\frac{n}{2}} \int_{\Delta_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^n d\mathbf{u},$$

where we have enlarged the domain of the integration on the right hand side to Δ_{k-1} , and the constant C_k is given by

$$C_k = \frac{|\Delta_{k-1} \cap E_{k-1}^c|}{|\Delta_{k-1} \cap \mathcal{E}_{k-1}|}.$$

To estimate C_k , we note that for all sufficiently large k , we have $|\Delta_{k-1} \cap E_{k-1}^c| \geq \frac{1}{2} |\Delta_{k-1}|$ and the components of $\mathbf{v} \in \mathcal{E}_{k-1}$ have disjoint domains. Thus, we get

$$C_k \leq C(4ek)^k.$$

The lemma then follows immediately by taking $n = k^2 \log k$. \square

Remark 3.4. In fact, $n \sim k \log k$ would be sufficient for the above proposition to be true. We take n much larger than that value in order to get good stability properties for the reconstruction procedure in Section 4.

So far, we have proved that the integral (3.1) is concentrated near the points $\{\frac{j}{k}\}$. As a consequence, the sum of the symmetrized signatures cannot be too far away from its maximal possible value as $n \rightarrow +\infty$. This is the content of the following theorem.

Theorem 3.5. *For all large enough k and $n = k^2 \log k$, there exists a word $|w^*| = k - 1$ such that*

$$\sum_{\ell} |\mathcal{S}_k^{2n}(w^*, \ell)| \geq \left(\frac{L}{7}\right)^k \int_{\Delta_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n} d\mathbf{u},$$

where the sum is taken over all multi-indices $\ell = (\ell_1, \dots, \ell_k)$ whose k components all run over $0, 1, \dots, n$.

Proof. For any word $|w| = k - 1$, by the expression (2.3), summing over the multi-indices ℓ gives

$$\sum_{\ell} |\mathcal{S}_k^{2n}(w, \ell)| \geq \left| \int_{\Delta_{k-1}} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \prod_{j=1}^k \sum_{\ell} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} d\mathbf{u} \right|, \quad (3.5)$$

where we have interchanged the sum over ℓ and the product over j since different components of ℓ are summed up independently. The integrand of the right hand

side of (3.5) can be split into two products: the pointwise derivatives $\dot{\gamma}_{u_j}^{i_j}$ and the increments $\sum_{\ell} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j}$. For the latter one, since

$$\sum_{\ell} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} = \frac{1}{2} \left((\Delta_{u_j} x + \Delta_{u_j} y)^{2n} + (\Delta_{u_j} x - \Delta_{u_j} y)^{2n} \right),$$

which is bounded above by $|\Delta_{u_j} \gamma|^{2n}$ and bounded below by $\frac{1}{2} |\Delta_{u_j} \gamma|^{2n}$ (recall the definition of the increments $|\Delta_{u_j} \gamma|$ from (2.1)), we have

$$\frac{1}{2^k} \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n} \leq \prod_{j=1}^k \sum_{\ell} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} \leq \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n}. \quad (3.6)$$

Now we look at the first part, namely $\prod_j \dot{\gamma}_{u_j}^{i_j}$. Since we expect the whole integral on the right hand side of (3.5) to be concentrated in the domain E_{k-1} , we choose a word $w^* = e_{i_1} * \dots * e_{i_{k-1}}$ such that

$$|\dot{\gamma}_{u_j}^{i_j}| \geq \frac{L}{3} \quad (3.7)$$

for all j and all $\mathbf{u} \in \Delta_{k-1} \cap E_{k-1}$, and that none of the $\dot{\gamma}_{u_j}^{i_j}$ changes its sign in the domain. The continuity of $\dot{\gamma}$ ensures that we can always find such a word as long as k is large enough. The main purpose of choosing w in this way is to ensure that the term $\prod_j \dot{\gamma}_{u_j}^{i_j}$ does not cause any degeneracy or cancellations in the integration in the domain of concentration $\Delta_{k-1} \cap E_{k-1}$.

We now decompose the right hand side of (3.5) into integrals over two disjoint domains: $\Delta_{k-1} \cap E_{k-1}$ and $\Delta_{k-1} \cap E_{k-1}^c$. For the first one, since the product $\prod_j \dot{\gamma}_{u_j}^{i_j}$ is bounded away from 0 by $(L/3)^{k-1}$ and does not change its sign in E_{k-1} , and the rest of the integrand is always positive as it only contains even powers, we can move the absolute value into the integral and combine (3.6) and (3.7) to get

$$\begin{aligned} & \left| \int_{\Delta_{k-1} \cap E_{k-1}} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \prod_{j=1}^k \sum_{\ell} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} d\mathbf{u} \right| \\ & \geq \left(\frac{L}{6} \right)^{k-1} \int_{\Delta_{k-1} \cap E_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n} d\mathbf{u}. \end{aligned}$$

On the other hand, it follows from the second inequality in (3.6) and Proposition 3.3 that the integration over the domain $\Delta_{k-1} \cap E_{k-1}^c$ is bounded by

$$e^{-2cn} \int_{\Delta_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n} d\mathbf{u}$$

for all large k and $n = k^2 \log k$. Thus, we obtain

$$\sum_{\ell} |\mathcal{S}_k^{2n}(w^*, \ell)| \geq c_k \int_{\Delta_{k-1} \cap E_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n} d\mathbf{u} - e^{-2cn} \int_{\Delta_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n} d\mathbf{u}, \quad (3.8)$$

where c_k could be chosen to be $(L/6)^{k-1}$. Applying Proposition 3.3 for another time, we can enlarge the domain of integration on the right hand side of (3.8) to Δ_{k-1} at the cost of having a slightly smaller constant c_k . This completes the proof of the theorem. \square

4 Reconstructing the path

We are now ready to reconstruct the path from its signature. Recall that our aim is to find the parameters $\rho_j \in [0, 1]$, $a_x^{(j)}, a_y^{(j)} \in \{0, 1\}$ and $\tilde{L} \in \mathbb{R}^+$ such that

$$\sup_{0 \leq j \leq k} \sup_{u \in [\frac{j-1}{k}, \frac{j}{k}]} |\tilde{L}(a_x^{(j)} \rho_j, a_y^{(j)}(1 - \rho_j)) - \dot{\gamma}_u| < C\eta_k,$$

where η_k is given by $\eta_k = \delta_{3/\epsilon_k}$, and ϵ_k is introduced in (3.4). We will recover these parameters in the order of ρ_j 's, $a_x^{(j)}, a_y^{(j)}$'s, and then finally \tilde{L} .

Remark 4.1. As for the magnitude of η_k , since $\epsilon_k \gtrsim \frac{1}{\sqrt{k}}$, it follows immediately that

$$\frac{1}{\sqrt{k}} \lesssim \eta_k \ll 1.$$

If $\dot{\gamma} \in \mathcal{C}^\alpha$ so that $\delta_k \sim k^{-\alpha}$, we would have $\eta_k \sim k^{-\frac{\alpha^2}{2}}$.

4.1 The unsigned directions

We start with the recovery of unsigned directions ρ_j 's of each piece. At this stage, we are only using the quantities $\mathcal{S}_k^{2n}(w, \ell)$ which can be obtained from the symmetrization and has an expression as in (2.3). Since we expect ρ_j to be close to the increment of the j -th piece of γ , it is natural to introduce for each j the unique real number $r_j \in [0, 1]$ such that

$$|\Delta_j x| : |\Delta_j y| = r_j : (1 - r_j).$$

We also let ℓ_{-j} be

$$\ell_{-j} = (\ell_1, \dots, \ell_{j-1}, \ell_{j+1}, \dots, \ell),$$

and write $\ell = (\ell_j, \ell_{-j})$. The following lemma will be useful in sequel.

Lemma 4.2. *For any $\mathbf{u} \in \Delta_{k-1} \cap E_{k-1}$, we have*

$$\sup_j \left| \frac{|\Delta_{u_j} x|}{|\Delta_{u_j} \gamma|} - r_j \right| < \frac{\eta_k}{L}.$$

Proof. First fix $\mathbf{u} \in \Delta_{k-1}$ and $0 \leq j \leq k$, and write

$$Q_j = \left[\frac{j-1}{k}, \frac{j}{k} \right] \cup [u_{j-1}, u_j].$$

We may assume without loss of generality that $r_j \geq \frac{1}{2}$, for otherwise we could get the same bound through

$$\left| \frac{|\Delta_{u_j} x|}{|\Delta_{u_j} \gamma|} - r_j \right| = \left| \frac{|\Delta_{u_j} y|}{|\Delta_{u_j} \gamma|} - (1 - r_j) \right|.$$

Now, since k is large enough, $r_j \geq \frac{1}{2}$ implies that

$$\inf_{t \in Q_j} |\dot{x}(t)| \geq \frac{L}{3}.$$

In particular, x is monotone in Q_j and \dot{x} does not change its sign. If y is also monotone in Q_j , then by the intermediate value theorem, there exist $v, \tilde{v} \in Q_j$ such that

$$\frac{|\Delta_{u_j} x|}{|\Delta_{u_j} \gamma|} = \frac{|\dot{x}(v)|}{L}, \quad r_j = \frac{|\dot{x}(\tilde{v})|}{L},$$

and the bound follows since $|Q_j| < 3\epsilon_k$. If y is not monotone in Q_j , then $\dot{y}(v) = 0$ for some $v \in Q_j$, and the continuity of $\dot{\gamma}$ implies $|\dot{y}(v)| \leq \delta_{\frac{3}{\epsilon_k}}$, which in turn gives

$$|\dot{x}(v)| \geq L - \delta_{\frac{3}{\epsilon_k}}, \quad \forall v \in Q_j.$$

This implies that both $\left| \frac{|\Delta_{u_j} x|}{|\Delta_{u_j} \gamma|} - r_j \right|$ and r_j are bounded below by $\frac{L - \delta_{\frac{2}{\epsilon_k}}}{L}$ and bounded above by 1, and the desired bound follows. \square

We are now ready to prove the following main theorem about the unsigned directions.

Theorem 4.3. *For all sufficiently large k and $n = k^2 \log k$, we have*

$$\sup_{0 \leq j \leq k} \left(\sum_w \sum_{\left| \frac{\ell_j}{n} - r_j \right| \geq 2\eta_k} \sum_{\ell-j} |\mathcal{S}_k^{2n}(w, \ell)| \right) / \left(\sum_w \sum_{\ell} |\mathcal{S}_k^{2n}(w, \ell)| \right) < e^{-k}, \quad (4.1)$$

where the sum is taken over all words $|w| = k - 1$ and all multi-indices $\ell = (\ell_1, \dots, \ell_k)$ within the appropriate range as indicated above.

Proof. Since the right hand side of the above bound does not depend on j , we only need to prove the bound for any fixed j . Also, as both the numerator and denominator in (4.1) scale like L^{2n+k-1} , we can assume without loss of generality that $L = 1$. By Theorem 3.5, the denominator satisfies

$$\sum_w \sum_{\ell} |\mathcal{S}_k^{2n}(w, \ell)| \geq \left(\frac{1}{7} \right)^{k-1} \int_{\Delta_{k-1}} \prod_{j=1}^k |\Delta_{u_j} \gamma|^{2n} d\mathbf{u}. \quad (4.2)$$

As for the numerator, since $L = 1$, we necessarily have $|\dot{x}_u|, |\dot{y}_u| \leq 1$. It then follows from the expression (2.3) that for each word w and multi-index ℓ , we have

$$|\mathcal{S}_k^{2n}(w, \ell)| \leq \int_{\Delta_{k-1}} \prod_{j=1}^k \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} d\mathbf{u}.$$

Now fixing w and j , and summing over ℓ in the appropriate ranges, we get

$$\begin{aligned} & \sum_{\left| \frac{\ell_j}{n} - r_j \right| > \epsilon_k} \sum_{\ell-j} |\mathcal{S}_k^{2n}(w, \ell)| \\ & < \frac{1}{2} \int_{\Delta_{k-1} \cup E_{k-1}} \sum_{\left| \frac{\ell_j}{n} - r_j \right| > \epsilon_k} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} \prod_{i \neq j} |\Delta_{u_i} \gamma|^{2n} d\mathbf{u}, \end{aligned}$$

where we have applied Proposition 3.3 to restrict the domain of integration to $\Delta_{k-1} \cup E_{k-1}$.

If the sum of ℓ_j were over the whole range $0 \leq \ell_j \leq n$, then the whole integrand of the right hand side above would just have the same order as $\prod_{j=1}^k |\Delta_{u_j}|^{2n}$. But since the quantity

$$\binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} \quad (4.3)$$

is concentrated near the values of ℓ_j such that $\frac{\ell_j}{n} \sim \frac{|\Delta_{u_j} x|}{|\Delta_{u_j} y|}$, and the range of the sum is away from this concentration mass, it is reasonable to expect that the whole sum will be very small as k and n grows.

In fact, by Lemma 4.2, if $\mathbf{u} \in \Delta_{k-1} \cap E_{k-1}$ and $|\frac{\ell_j}{n} - r_j| > 2\eta_k$, we will have

$$|\frac{\ell_j}{n} - p| > \eta_k,$$

where $p = \frac{|\Delta_{u_j} x|}{|\Delta_{u_j} y|}$. Since quantities in (4.3) approximate $\mathcal{N}(2np, 2np(1-p))$ distribution with

$$\frac{p}{1-p} = \frac{|\Delta_{u_j} x|}{|\Delta_{u_j} y|}$$

when ℓ_j ranges from 0 to n , it follows from standard Gaussian integration that

$$\sum_{|\frac{\ell_j}{n} - r_j| \geq 2\eta_k} \sum_{\ell-j} |\mathcal{S}_k^{2n}(w, \ell)| < \sum_{|\frac{\ell_j}{n} - p| \geq \eta_k} \sum_{\ell-j} |\mathcal{S}_k^{2n}(w, \ell)| < e^{-n\eta_k^2}.$$

Since the above bound is true for every word w , and there are totally 2^{k-1} choices of w 's, we then obtain the bound for the numerator as

$$\sum_w \sum_{|\frac{\ell_j}{n} - r_j| > 2\eta_k} \sum_{\ell-j} |\mathcal{S}_k^{2n}(w, \ell)| < 2^{k-1} e^{-n\eta_k^2}. \quad (4.4)$$

Recall from Remark 4.1 that $\eta_k \gtrsim \frac{1}{\sqrt{k}}$, the claim then follows by combining (4.2) and (4.4) with $n = k^2 \log k$. \square

The following easy corollary enables one to select the directions ρ_j (up to the sign) for each piece $\tilde{\gamma}_j$.

Corollary 4.4. [*Choosing directions*] *Let k be a fixed large number. For each $0 \leq j \leq k$, there exists $\rho_j \in [0, 1]$ such that*

$$\left(\sum_w \sum_{j=1}^k \sum_{|\frac{\ell_j}{n} - \rho_j| < 2\eta_k} |\mathcal{S}_k^{2n}(w, \ell)| \right) / \left(\sum_w \sum_{\ell} |\mathcal{S}_k^{2n}(w, \ell)| \right) > \frac{1}{2}. \quad (4.5)$$

Moreover, if $\{\rho_j\}$ is any set that satisfies (4.5), then we must have

$$|\rho_j - r_j| < 3\eta_k$$

for all $j = 1, \dots, k$.

Proof. The existence of the set $\{\rho_j\}$ that satisfies (4.5) follows directly by setting $\rho_j = r_j$ and applying Theorem 4.3. On the other hand, if $|\rho_j - r_j| \geq 3\eta_k$ for some j , then $|\frac{\ell_j}{n} - \rho_j| < \eta_k$ implies that

$$|\frac{\ell_j}{n} - r_j| \geq 2\eta_k.$$

By Theorem 4.3, this set of $\{\rho_j\}$ must violate (4.5). This completes the proof. \square

Remark 4.5. Note that the left hand side of (4.5) are all information that is available from the signature X . If we choose $\{\rho_j\}$ according to (4.5), the above corollary tells that these unsigned directions we recover from the signature must be close to the true directions $\{r_j\}$.

Remark 4.6. The readers might have noticed that the criterion of choosing the ρ_j 's above involve the knowledge of η_k . This is of course not a problem if we know the modulus of continuity of $\dot{\gamma}$ in advance. But even if that information is not available, since $\dot{\gamma}$ is continuous, one could always use a sequence $\{\eta_k\}$ that goes to 0 as slow as possible. This will still produce the asymptotically correct directions $\{\rho_j\}$, though at the cost of a larger error η_k .

4.2 The signs

We now turn to the recovery of the sign of the direction of each piece. For this, we need to move one level up in the signatures, which requires extra information than $\mathcal{S}_k^{2n}(w, \ell)$'s.

We now look at the signature at level $2nk + k$ (in addition to the level $2nk + k - 1$ before), and divide into k blocks of size $2n$ except one of them which has size $2n + 1$, still with one letter separating consecutive blocks.

More precisely, for any word $w = e_{i_1} * \dots * e_{i_{k-1}}$, any multi-index $\ell = (\ell_1, \dots, \ell_k)$ with $0 \leq \ell_j \leq n$, and any $1 \leq i \leq k$, we let $\mathcal{W}_{k,i,x}^{2n}(w, \ell)$ denote the set of words

$$w' = w_1 * e_{i_1} * \dots * e_{i_{k-1}} * w_k$$

such that $|w_j(x)| = 2\ell_j$ for each $j \neq i$ while $|w_i(x)| = 2\ell_i + 1$. The set $\mathcal{W}_{k,i,x}^{2n}(w, \ell)$ is different from $\mathcal{W}_k^{2n}(w, \ell)$ as in Section 3 in that the i -th block has size $2n + 1$ instead of $2n$, and contains $2\ell_i + 1$ x 's instead of $2\ell_i$.

We then define the quantity $\mathcal{S}_{k,i,x}^{2n}(w, \ell)$ to be

$$\mathcal{S}_{k,i,x}^{2n}(w, \ell) = \sum_{w' \in \mathcal{W}_{k,i,x}^{2n}(w, \ell)} C(w'). \quad (4.6)$$

The aim of introducing this quantity is to recover the sign of x -direction in the i -th piece of the path via comparison with $\mathcal{S}_k^{2n}(w, \ell)$. Similarly, we define the quantity $\mathcal{S}_{k,i,y}^{2n}(w, \ell)$ to be the sum which is the same as above except that the $|w_i(y)| = 2n - 2\ell_i + 1$. It is not hard to check that these quantities can be expressed as

$$\begin{aligned} \mathcal{S}_{k,i,x}^{2n}(w, \ell) = & \int_{\Delta_{k-1}} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \cdot \binom{2n+1}{2\ell_i+1} (\Delta_{u_i} x)^{2\ell_i+1} (\Delta_{u_i} y)^{2n-2\ell_i} \\ & \prod_{j \neq i} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} d\mathbf{u}, \end{aligned} \quad (4.7)$$

and

$$\mathcal{S}_{k,i,y}^{2n}(w, \ell) = \int_{\Delta_{k-1}} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \cdot \binom{2n+1}{2\ell_i} (\Delta_{u_i} x)^{2\ell_i} (\Delta_{u_i} y)^{2n+1-2\ell_i} \prod_{j \neq i} \binom{2n}{2\ell_j} (\Delta_{u_j} x)^{2\ell_j} (\Delta_{u_j} y)^{2n-2\ell_j} d\mathbf{u}.$$

Again, (4.6) is the information available from the signature, and (4.7) is an expression for this quantity. The same is true for $\mathcal{S}_{k,i,y}^{2n}(w, \ell)$.

In order to use the quantities (4.6) to determine the signs, we first choose a word w^* so that $\prod_j \dot{\gamma}_{u_j}^{i_j}$ is non-degenerate in the region $\Delta_{k-1} \cap E_{k-1}$. This can be achieved as follows:

Let $\{\rho_j\}$ be the set of unsigned directions chosen according to Corollary 4.4, then for each $j = 1, \dots, k-1$ we let

$$\begin{aligned} e_{i_j} &= x, & \text{if } \rho_j &\geq \frac{1}{2}, \\ e_{i_j} &= y, & \text{if } \rho_j &< \frac{1}{2}, \end{aligned} \tag{4.8}$$

and we choose the word w^* to be

$$w^* = e_{i_1} * \dots * e_{i_{k-1}}. \tag{4.9}$$

In fact, for large enough k , by Corollary 4.4, this choice of w^* necessarily guarantees that

$$\prod_{j=1}^{k-1} |\dot{\gamma}_{u_j}^{i_j}| \geq \left(\frac{L}{3}\right)^{k-1}$$

for all $\mathbf{u} \in E_{k-1}$ and that the product does not change its sign in this domain. Note that Theorem 3.5 only gives the existence of such a word w^* while here we choose it explicitly based on the recovery of the unsigned directions. We now determine the signs of the i -th piece (depending on n) as follows.

Definition 4.7. [Determining the signs] Fix k and let $n = k^2 \log k$. Let w^* be the word chosen by (4.8) and (4.9). We then choose the sign $a_x^{(i)}, a_y^{(i)} \in \{\pm 1\}$ for the i -th linear piece by:

$$\begin{aligned} a_x^{(i)} &= 1, & \text{if } \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,i,x}^{2n}(w^*, \ell)} &\geq 0, \\ a_x^{(i)} &= -1, & \text{if } \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,i,x}^{2n}(w^*, \ell)} &< 0. \end{aligned}$$

The choice for $a_y^{(i)}$ is the same except replacing $\mathcal{S}_{k,i,x}^{2n}(w^*, \ell)$ by $\mathcal{S}_{k,i,y}^{2n}(w^*, \ell)$.

It appears that the above choices of the signs depend on k , and the choices may be different if k changes. But it turns out that the choices above remain stable for all sufficiently large k , and they indeed give the correct signs as long as the directions are not close to degenerate. This is the content of the following theorem.

Theorem 4.8. *Let $n = k^2 \log k$. If $r_i \geq 3\epsilon_k$, then*

$$\begin{aligned} \liminf_{k \rightarrow +\infty} k^k \cdot \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,i,x}^{2n}(w^*, \ell)} &= +\infty & \text{if } \Delta_i x > 0, \\ \limsup_{k \rightarrow +\infty} k^k \cdot \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,i,x}^{2n}(w^*, \ell)} &= -\infty & \text{if } \Delta_i x < 0. \end{aligned}$$

Similarly, if $r_i \leq 1 - 3\epsilon_k$, then

$$\begin{aligned} \liminf_{k \rightarrow +\infty} k^k \cdot \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,i,y}^{2n}(w^*, \ell)} &= +\infty & \text{if } \Delta_i y > 0, \\ \limsup_{k \rightarrow +\infty} k^k \cdot \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,i,y}^{2n}(w^*, \ell)} &= -\infty & \text{if } \Delta_i y < 0. \end{aligned}$$

Remark 4.9. The above theorem guarantees that as long as the i -th piece of γ is not too horizontal or vertical (corresponding to the assumptions $r_i \geq 3\epsilon_k$ and $r_i \leq 1 - 3\epsilon_k$), then the choice in Definition 4.7 do give the correct signs for all sufficiently large k . The case $r_i < 3\epsilon_k$ or $r_i > 1 - 3\epsilon_k$ is not covered, but since the i -th piece would be almost horizontal or vertical in that situation, the choice of the sign would not affect accuracy.

Proof. We only prove the first case when $r_i \geq 3\epsilon_k$ and $\Delta_i x > 0$; the rest are essentially the same. For each $\mathbf{u} \in \Delta_{k-1}$, write

$$\mathcal{N}(\mathbf{u}) = \frac{1}{2^k} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \prod_{j=1}^k \left((\Delta_{u_j} x + \Delta_{u_j} y)^{2n} + (\Delta_{u_j} x - \Delta_{u_j} y)^{2n} \right),$$

and

$$\mathcal{D}(\mathbf{u}) = \frac{1}{2^k} \prod_{j=1}^{k-1} \dot{\gamma}_{u_j}^{i_j} \prod_{j=1}^k \left((\Delta_{u_j} x + \Delta_{u_j} y)^{2n+\delta_{i,j}} + (\Delta_{u_j} x - \Delta_{u_j} y)^{2n+\delta_{i,j}} \right),$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. By the expressions (2.3) and (4.7), the numerator and denominator could then be written as

$$\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell) = \int_{\Delta_{k-1}} \mathcal{D}(\mathbf{u}) d\mathbf{u}, \quad \sum_{\ell} \mathcal{S}_{k,i,x}^{2n}(w^*, \ell) = \int_{\Delta_{k-1}} \mathcal{N}(\mathbf{u}) d\mathbf{u},$$

and we need to estimate the ratio of the two integrals. Same as before, by Proposition 3.3, we only need to consider integration over the domain $\Delta_{k-1} \cap E_{k-1}$, and we hope to replace the ratio of integrals by a pointwise bound on

$$\frac{\mathcal{N}(\mathbf{u})}{\mathcal{D}(\mathbf{u})} = \frac{(\Delta_{u_i} x + \Delta_{u_i} y)^{2n} + (\Delta_{u_i} x - \Delta_{u_i} y)^{2n}}{(\Delta_{u_i} x + \Delta_{u_i} y)^{2n+1} + (\Delta_{u_i} x - \Delta_{u_i} y)^{2n+1}}.$$

For this, we first note that the assumptions $r_i \geq 3\epsilon_k$ and $\Delta_i x > 0$ imply that $\Delta_{u_i} x > 0$ for all $\mathbf{u} \in \Delta_{k-1} \cap E_{k-1}$. Thus, the ratio $\frac{\mathcal{N}(\mathbf{u})}{\mathcal{D}(\mathbf{u})}$ is always positive in $\Delta_{k-1} \cap E_{k-1}$. Since

$$\max \{ |\Delta_{u_i} x + \Delta_{u_i} y|, |\Delta_{u_i} x - \Delta_{u_i} y| \} = |\Delta_{u_i} \gamma|,$$

we then have the bound

$$\frac{\mathcal{N}(\mathbf{u})}{\mathcal{D}(\mathbf{u})} \geq \frac{1}{|\Delta_{u_i} \gamma|} \geq \frac{1}{3\epsilon_k L},$$

uniformly in $\mathbf{u} \in \Delta_{k-1} \cap E_{k-1}$. Thus, we obtain

$$\frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,i,x}^{2n}(w^*, \ell)} \geq \frac{1}{3\epsilon_k L} |\Delta_{k-1} \cap E_{k-1}|.$$

The theorem then follows by noting that $\epsilon_k \rightarrow 0$ and $|\Delta_{k-1} \cap E_{k-1}| > (1/k)^k$. \square

4.3 Length

We have now obtained for each j the signed direction $(a_x^{(j)} \rho_j, a_y^{(j)} (1 - \rho_j))$, and the only remaining quantity to be determined is \tilde{L} , which is expected to approximate the ℓ^1 length of γ . We can achieve this by a simple scaling argument. In fact, if $|\gamma|_{\ell^1} = L$, then by Corollary 4.4, Theorem 4.8, and the regularity of $\dot{\gamma}$, we will have

$$\sup_{0 \leq j \leq k} \sup_{u \in [\frac{j-1}{k}, \frac{j}{k}]} \left| L(a_x^{(j)} \rho_j, a_y^{(j)} (1 - \rho_j)) - \dot{\gamma}_u \right| < C\eta_k$$

where C depends on the path γ but not k . In particular, this implies (by integrating u from 0 to 1)

$$|X^1| - C\eta_k < L \left(\left| \sum_j a_x^{(j)} \rho_j \right| + \left| \sum_j a_y^{(j)} (1 - \rho_j) \right| \right) < |X^1(\gamma)| + C\eta_k, \quad (4.10)$$

where $|X^1| = |X^1(\gamma)| = |x_1| + |y_1|$ is the ℓ^1 norm of the increment.

If $|X^1| > 0$, then η_k becomes negligible with respect to $|X^1|$ for large k , and (4.10) guarantees that the quantity $\left| \sum_j a_x^{(j)} \rho_j \right| + \left| \sum_j a_y^{(j)} (1 - \rho_j) \right|$ is strictly positive and bounded away from 0 uniformly in k . Then, (4.10) suggests that it is natural to set

$$\tilde{L} := \frac{|X^1|}{\left| \sum_j a_x^{(j)} \rho_j \right| + \left| \sum_j a_y^{(j)} (1 - \rho_j) \right|}, \quad (4.11)$$

and clearly this choice of \tilde{L} satisfies

$$\sup_{u \in [\frac{j-1}{k}, \frac{j}{k}]} \left| \tilde{L}(a_x^{(j)} \rho_j, a_y^{(j)} (1 - \rho_j)) - \dot{\gamma}_u \right| < C\eta_k. \quad (4.12)$$

In the case $|X^1| = 0$, one could not simply neglect η_k as $\left| \sum_j a_x^{(j)} \rho_j \right| + \left| \sum_j a_y^{(j)} (1 - \rho_j) \right|$ would also be close to 0. The expression (4.11) determining \tilde{L} would then have the form of $\frac{0}{0}$, which causes a problem of the definition. The way we circumvent it is to attach a linear piece of positive length to the end of γ . More precisely, we define

$$Y := X(\gamma) \otimes \exp(a_x^{(k)} \rho_k x + a_y^{(k)} (1 - \rho_k) y),$$

where $\rho_k, a_x^{(k)}$ and $a_y^{(k)}$ are the relevant coefficients of the last linear piece from the reconstruction. Then, Y is the signature of $\gamma * \beta$ where

$$\beta = a_x^{(k)} \rho_k x + a_y^{(k)} (1 - \rho_k) y$$

is a line segment of length 1. The choices of $a_x^{(k)}, a_y^{(k)}$ and ρ_k ensures that β concatenates almost smoothly to the end of γ . In particular, it will not create any cancellations to the original path γ . It is clear that

$$|Y^1| = 1 \neq 0,$$

and we can apply the previous procedure to the new signature Y to get a path asymptotically close to $\gamma * \beta$. Finally, removing β from that path recovers γ . This finishes the choice of length as well as the whole reconstruction procedure.

4.4 Summary

We now end this section by summarizing the symmetrization procedure as follows.

1. Let k be a fixed large number. For each $j = 1, \dots, k-1$, choose $\rho_j \in [0, 1]$ according to Corollary 4.4 such that

$$\left(\sum_w \sum_{j=1}^k \sum_{|\frac{\ell_j}{n} - \rho_j| < 2\eta_k} |\mathcal{S}_k^{2n}(w, \ell)| \right) / \left(\sum_w \sum_{\ell} |\mathcal{S}_k^{2n}(w, \ell)| \right) > \frac{1}{2},$$

and any set $\{\rho_j\}$ satisfying the above would suffice.

2. Now, we choose the word $w^* = e_{i_1} * \dots * e_{i_{k-1}}$ by

$$\begin{aligned} e_{i_j} &= x, & \text{if } \rho_j \geq \frac{1}{2}; \\ e_{i_j} &= y, & \text{if } \rho_j < \frac{1}{2}, \end{aligned}$$

and determine the signs $a_x^{(j)}$ and $a_y^{(j)}$ by

$$\begin{aligned} a_x^{(j)} &= 1, & \text{if } \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,j,x}^{2n}(w^*, \ell)} \geq 0; \\ a_x^{(j)} &= -1, & \text{if } \frac{\sum_{\ell} \mathcal{S}_k^{2n}(w^*, \ell)}{\sum_{\ell} \mathcal{S}_{k,j,x}^{2n}(w^*, \ell)} < 0, \end{aligned}$$

where w^* is the word chosen above.

3. Finally, we determine the length \tilde{L} by

$$\tilde{L} := \frac{|X^1|}{|\sum_j a_x^{(j)} \rho_j| + |\sum_j a_y^{(j)} (1 - \rho_j)|}$$

if $|X^1| \neq 0$. In the case $|X^1| = 0$, we recovery \tilde{L} and the path γ following the procedure in the second half of Section 4.3.

Corollary 4.4 and Theorem 4.8 guarantee that we have the following theorem.

Theorem 4.10. *Let k be a large integer, and $\{\rho_j\}, \{a_x^{(j)}\}, \{a_y^{(j)}\}$ and \tilde{L} be determined in the above procedure. Then, we have*

$$\sup_{0 \leq j \leq k} \sup_{u \in [\frac{j-1}{k}, \frac{j}{k}]} \left| \tilde{L}(a_x^{(j)} \rho_j, a_y^{(j)} (1 - \rho_j)) - \dot{\gamma}_u \right| < C \eta_k.$$

In particular, if we let

$$\tilde{\gamma}_j = a_x^{(j)} \rho_j x + a_y^{(j)} \rho_j y, \quad (4.13)$$

and $\tilde{\gamma} = \tilde{\gamma}_1 * \dots * \tilde{\gamma}_k$, then $\|\tilde{\gamma} - \gamma\|_{\mathcal{C}^1} < C \eta_k$ when both are parametrized at uniform speed.

Remark 4.11. The path $\tilde{\gamma}$ produced from this symmetrization procedure is not actually not unique, as there are certainly infinitely many sets $\{\rho_j\}$ that satisfy (4.5). In fact, what this procedure produces is not a single path, but instead a measure on piecewise linear paths which converges to the delta measure on the original path γ if $\gamma \in \mathcal{C}^1$. In practice, one can just choose any piecewise linear path from the area where the measure is concentrated, and this path must be close to γ in 1-variation with the explicit error bound (4.13). In fact, this is what we have done in the above formulation.

4.5 Higher dimensions

We now briefly explain how the procedure developed above carries to paths in d dimensions in essentially the same way. In this case, each linear piece $\tilde{\gamma}_j$ has the form

$$\tilde{\gamma}_j = \frac{\tilde{L}}{k} \left(a_1^{(j)} \rho_1^{(j)} e_1 + \dots + a_d^{(j)} \rho_d^{(j)} e_d \right),$$

where $(\rho_1^{(j)}, \dots, \rho_d^{(j)})$ is a non-negative vector with $\sum_i \rho_i^{(j)} = 1$ representing the unsigned direction of the j -th piece, $a_i^{(j)} \in \{\pm 1\}$ represents the signs of each direction, and \tilde{L} approximates the ℓ^1 length of the path.

The symmetrization procedure as well as the concentration properties are the same as in the two-dimensional case, except that each ℓ_j now itself is a vector. This then gives the difference in recovering the unsigned directions, where one needs to replace the binomial argument by a multinomial argument. As soon as one gets the directions accurately, the recovery of the signs as well as the length then follows immediately in the same way.

References

- [BG14] H.Boedihardjo, X.Geng, On the uniqueness of signature problem through a strengthened Le Jan-Qian approximation scheme, Arxiv preprint: 1401.6165, 2014.
- [BGLY14] H.Boedihardjo, X.Geng, T.J.Lyons, D.Yang, The signature of a rough path: uniqueness, Arxiv preprint: 1406.7871.

- [Che54] K.-T.Chen, Iterated integrals and exponential homomorphisms, *Proceedings of the London Mathematical Society*, Vol.4, No.3 (1954), 502-512.
- [Che57] K.-T.Chen, Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula, *Annals of Mathematics*, Vol.65, No.1 (1957), 163-178.
- [Che58] K.-T.Chen, Integration of paths - a faithful representation of paths by noncommutative formal power series, *Transactions of the A.M.S.*, Vol.89, No.2 (1958), 395-407.
- [Che77] K.-T.Chen, Iterated path integrals, *Bulletin of the A.M.S.*, Vol.83, No.5 (1977), 831-879.
- [GQ15] X.Geng, Z.Qian, On an inversion theorem for Stratonovich's signatures of multidimensional diffusion paths, to appear in *Annales de l'Institut Henri Poincaré*, 2015.
- [HL08] B.M.Hambly, T.J.Lyons, Some notes on trees and paths, unpublished manuscript, available at <http://arxiv.org/abs/0809.1365>, 2008.
- [HL10] B.M.Hambly, T.J.Lyons, Uniqueness for the signature of a path of bounded variation and the reduced path group, *Annals of Mathematics*, Vol.171, No.1 (2010), 109-167.
- [LJQ13] Y.Le Jan, Z.Qian, Stratonovich's signatures of Brownian motion determine Brownian sample paths, *Probability Theory and Related Fields*, Vol.157, Issue 1-2 (2013), 209-223.
- [Lyo98] T.J.Lyons, Differential equations driven by rough signals, *Rev. Mat. Iberoamericana*, Vol.14, No.2 (1998), 215-310.
- [LX15] T.J.Lyons, W.Xu, Hyperbolic development and inversion of signature, Arxiv preprint, 2015.

MATHEMATICAL AND OXFORD-MAN INSTITUTES, UNIVERSITY OF OXFORD,
WOODSTOCK ROAD, OXFORD, OX2 6GG, UK.

Email: tlyons@maths.ox.ac.uk

MATHEMATICS INSTITUTE, UNIVERSITY OF WARWICK, COVENTRY, CV4 7AL,
UK.

Email: weijun.xu@warwick.ac.uk